UCRL-JC-122454 Abs

A FLEXIBLE DATABASE SCHEMA FOR LARGE-SCALE PHYSICAL MAPPING AND SEQUENCING ACROSS MULTIPLE GENOMES.

Tom Slezak (slezak@llnl.gov), Mark Wagner, T. Mimi Yeh. Human Genome Center, Biology and Biotechnology Research Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

The Human Genome Center at LLNL has developed a relational database over the last 6 years to support our work on chromosome 19. We intentionally designed the database to support our specific physical mapping needs. We anticipated applying what we learned form this experience when we expanded our capabilities to support new approaches and to cover other genetic real estate.

Effective closure of the physical map of ch19 has been declared and we are now retargeting our efforts towards production sequencing (and associated high-resolution mapping) of certain gene families across the entire human genome, regions of interest in other genomes with conserved synteny with respect to humans, and potential work on various bacterial, plant, and animal genomes. Separate databases for each chromosome or genome are not well-suited for the comparative biology that lies in our future. We must scale up our database to be able to handle queries on mapping and sequencing data that span all our targets regardless of species, provide for better public access to data, and fully participate in the Federation of Genome Databases.

At the last meeting we reported on our design ideas. Major concepts include: all objects have a unique, permanent identifier; objects and relations will be highly abstracted (single base-class tables for all clones, probes, hybridization results, etc.); each object and relation can be flagged as public or private; object storage handled separately, etc. Key to this design is a central global identifier table, which tracks information on every object and relation, providing a flexible and simple reference interface mechanism. This design stems from the generic "mappable object" abstraction which allowed us to successfully integrate our physical mapping data, as described at earlier meetings.

This database has been implemented and our 250Mb of ch19 data transferred to it, using a database re-engineering tool (described by Mark Wagner, et. al. in a separate poster.) We are currently implementing WWW interfaces to this database for data entry and non-graphical querying and have modified our graphical browser to read from it. We will discuss performance implications of this abstracted schema design from our early experience with it and extrapolate on its ability to scale to meet our future needs.

(This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.)